# Inference

# 9

## 9.1 Introduction

The first four chapters of this book illustrated how one can summarize a data set both numerically and graphically. The validity of interpretations made from such a descriptive analysis is valid only for the data set under consideration and cannot necessarily be generalized to other data. However, it is desirable to make conclusions about the entire population of interest and not only about the sample data. In this chapter, we describe the framework of **statistical inference** which allows us to infer from the sample data about the population of interest–at a given, prespecified uncertainty level–and knowledge about the random process generating the data.

Consider an example where the objective is to forecast an election outcome. This requires us to determine the proportion of votes that each of the $k$ participating parties is going to receive, i.e. to calculate or estimate $p_1, p_2, \ldots, p_k$. If it is possible to ask every voter about their party preference, then one can simply calculate the proportions $p_1, p_2, \ldots, p_k$ for each party. However, it is logistically impossible to ask all eligible voters (which form the population in this case) about their preferred party. It seems more realistic to ask only a small fraction of voters and infer from their responses to the responses of the whole population. It is evident that there might be differences in responses between the sample and the population—but the more voters are asked, the closer we are to the population's preference, i.e. the higher the precision of our estimates for $p_1, p_2, \ldots, p_k$ (the meaning of "precision" will become clearer later in this chapter). Also, it is intuitively clear that the sample must be a representative sample of the voters' population to avoid any discrepancy or bias in the forecasting. When we speak of a representative sample, we mean that all the characteristics present in the population are contained in the sample too. There are many ways to get representative random samples. In fact, there is a branch of statistics, called sampling theory, which studies this subject [see, e.g. Groves et al. (2009) or Kauermann and Küchenhoff (2011) for more details]. A simple random sample is one where each voter has an equal probability of being selected in the sample and



C. Heumann et al., *Introduction to Statistics and Data Analysis*,
DOI 10.1007/978-3-319-46162-5_9

each voter is independently chosen from the same population. In the following, we will assume that all samples are simple random samples. To further formalize the election forecast problem, assume that we are interested in the true proportions which each party receives on the election day. It is practically impossible to make a perfect prediction of these proportions because there are too many voters to interview, and moreover, a voter may possibly make their final decision possibly only when casting the vote and not before. The voter may change his/her opinion at any moment and may differ from what he/she claimed earlier. In statistics, we call these true proportions *parameters of the population*. The task is then to estimate these parameters on the basis of a sample. In the election example, the intuitive estimates for the proportions in the population are the proportions in the sample and we call them *sample estimates*. How to find good and precise estimates are some of the challenges that are addressed by the concept of *statistical inference*. Now, it is possible to describe the election forecast problem in a statistical and operational framework: estimate the parameters of a population by calculating the sample estimates. An important property of every good statistical inference procedure is that it provides not only estimates for the population parameters but also information about the precision of these estimates.

Consider another example in which we would like to study the distribution of weight of children in different age categories and get an understanding of the "normal" weight. Again, it is not possible to measure the weight of all the children of a specific age in the entire population of children in a particular country. Instead, we draw a random sample and use methods of statistical inference to estimate the weight of children in each age group. More specifically, we have several populations in this problem. We could consider all boys of a specific age and all girls of a specific age as two different populations. For example, all 3-year-old boys will form one possible population. Then, a random sample is drawn from this population. It is reasonable to assume that the distribution of the weight of $k$-year-old boys follows a normal distribution with some unknown parameters $\mu_{kb}$ and $\sigma^2_{kb}$. Similarly, another population of $k$-year-old girls is assumed to follow a normal distribution with some unknown parameters $\mu_{kg}$ and $\sigma^2_{kg}$. The indices $kb$ and $kg$ are used to emphasize that the parameters may vary by age and gender. The task is now to calculate the estimates of the unknown parameters (in the population) of the normal distributions from the samples. Using quantiles, a range of "normal" weights can then be specified, e.g. the interval from the 1 % quantile to the 99 % quantile of the estimated normal distribution or, alternatively, all weights which are not more than twice the standard deviation away from the mean. Children with weights outside this interval may be categorized as underweight or overweight. Note that we make a specific assumption for the distribution class; i.e. we assume a normal distribution for the weights and estimate its parameters. We call this a **parametric** estimation problem because it is based on distributional assumptions. Otherwise, if no distributional assumptions are made, we speak of a **nonparametric** estimation problem.

## 9.2 Properties of Point Estimators

As we discussed in the introduction, the primary goal in statistical inference is to find a good estimate of (a) population parameter(s). The parameters are associated with the probability distribution which is believed to characterize the population; e.g. $\mu$ and $\sigma^2$ are the parameters in a normal distribution $N(\mu, \sigma^2)$. If these parameters are known, then one can characterize the entire population. In practice, these parameters are unknown, so the objective is to estimate them. One can attempt to obtain them based on a function of the sample values. But what does this function look like; and if there is more than one such function, then which is the best one? What is the best approach to estimate the population parameters on the basis of a given sample of data? The answer is given by various statistical concepts such as bias, variability, consistency, efficiency, sufficiency, and completeness of the estimates. We are going to introduce them now.

Assume $x = (x_1, x_2 \ldots, x_n)$ are the observations of a random sample from a population of interest. The random sample represents the realized values of a random variable $X$. It can be said that $x_1, x_2 \ldots, x_n$ are the $n$ observations collected on the random variable $X$. Any function of random variables is called a **statistic**. For example, $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, $\max(X_1, X_2, \ldots, X_n)$ etc. are functions of $X_1, X_2, \ldots, X_n$, so they are a statistic. It follows that a statistic is also a random variable. Consider a statistic $T(X)$ which is used to estimate a population parameter $\theta$ (which may be either a scalar or a vector). We say $T(X)$ is an **estimator** of $\theta$. To indicate that we estimate $\theta$ using $T(X)$, we use the "hat" ( $\hat{}$ ) symbol, i.e. we write $\hat{\theta} = T(X)$. When $T$ is calculated from the sample values $x_1, x_2 \ldots, x_n$, we write $T(x)$ and call it an **estimate** of $\theta$. It becomes clear that $T(X)$ is a random variable but $T(x)$ is its observed value (dependent on the actual sample). For example, $T(X) = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is an estimator and a statistic, but $T(x) = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is its estimated value from the realized sample values $x_1, x_2, \ldots, x_n$. Since the sample values are realizations from a random variable, each sample leads to a different value of the estimate of the population parameter. The population parameter is assumed to be a fixed value. Parameters can also be assumed to be random, for example in Bayesian statistics, but this is beyond the scope of this book.

### 9.2.1 Unbiasedness and Efficiency

**Definition 9.2.1** An estimator $T(X)$ is called an *unbiased* estimator of $\theta$ if

$$\mathrm{E}_\theta(T(X)) = \theta\,. \tag{9.1}$$

The index $\theta$ denotes that the expectation is calculated with respect to the distribution whose parameter is $\theta$.

The bias of an estimator $T(X)$ is defined as

$$\text{Bias}_\theta(T(X)) = \text{E}_\theta(T(X)) - \theta \, . \tag{9.2}$$

It follows that an estimator is said to be unbiased if its bias is zero.

**Definition 9.2.2** The variance of $T(X)$ is defined as

$$\text{Var}_\theta(T(X)) = \text{E}\left\{[T(X) - \text{E}(T(X))]^2\right\} \, . \tag{9.3}$$

Both bias and variance are measures which characterize the properties of an estimator. In statistical theory, we search for "good" estimators in the sense that the bias and the variance are as small as possible and therefore the accuracy is as high as possible. Readers interested in a practical example may consult Examples 9.2.1 and 9.2.2, or the explanations for Fig. 9.1.

It turns out that we cannot minimize both measures simultaneously as there is always a so-called bias–variance tradeoff. A measure which combines bias and variance into one measure is the mean squared error.

**Definition 9.2.3** The mean squared error (MSE) of $T(X)$ is defined as

$$\text{MSE}_\theta(T(X)) = \text{E}\left\{[T(X) - \theta]^2\right\} \, . \tag{9.4}$$

The expression (9.4) can be partitioned into two parts: the variance and the squared bias, i.e.

$$\text{MSE}_\theta(T(X)) = \text{Var}_\theta(T(X)) + [\text{Bias}_\theta(T(X))]^2 \, . \tag{9.5}$$

This can be proven as follows:

$$\begin{aligned}
\text{MSE}_\theta(T(X)) &= \text{E}[T(X) - \theta]^2 \\
&= \text{E}[(T(X) - \text{E}_\theta(T(X)) + (\text{E}_\theta(T(X) - \theta)]^2 \\
&= \text{E}[T(X) - \text{E}_\theta(T(X))]^2 + [\text{E}_\theta(T(X)) - \theta]^2 \\
&= \text{Var}_\theta(T(X)) + [\text{Bias}_\theta(T(X))]^2 \, .
\end{aligned}$$

Note that the calculation is based on the result that the cross product term is zero. The mean squared error can be used to compare different biased estimators.

**Definition 9.2.4** An estimator $T_1(X)$ is said to be MSE-better than another estimator $T_2(X)$ for estimating $\theta$ if

$$\text{MSE}_\theta(T_1(X)) < \text{MSE}_\theta(T_2(X)) \, ,$$

where $\theta \in \Theta$ and $\Theta$ is the parameter space, i.e. the set of all possible values of $\theta$. Often, $\Theta$ is $\mathbb{R}$ or all positive real values $\mathbb{R}_+$. For example, for a normal distribution, $N(\mu, \sigma^2)$, $\mu$ can be any real value and $\sigma^2$ has to be a number greater than zero.

Unfortunately, we cannot find an MSE-optimal estimator in the sense that an estimator is MSE-better than all other possible estimators for all possible values of $\theta$. This becomes clear if we define the constant estimator $T(x) = c$ (independent of the actual sample): if $\theta = c$, i.e. if the constant value equals the true population parameter we want to estimate, then the MSE of this constant estimator is zero (but it will be greater than zero for all other values of $\theta$, and the bias increases more as we move $c$ far away from the true $\theta$). Usually, we can only find estimators which are locally best (in a certain subset of $\Theta$). This is why classical statistical inference restricts the search for best estimators to the class of unbiased estimators. For unbiased estimators, the MSE is equal to the variance of an estimator. In this context, the following definition is used for comparing two (unbiased) estimators.

**Definition 9.2.5** An unbiased estimator $T_1(X)$ is said to be more efficient than another unbiased estimator $T_2(X)$ for estimating $\theta$ if

$$\text{Var}_\theta(T_1(X)) \leq \text{Var}_\theta(T_2(X)) \ , \ \forall \theta \in \Theta \ ,$$

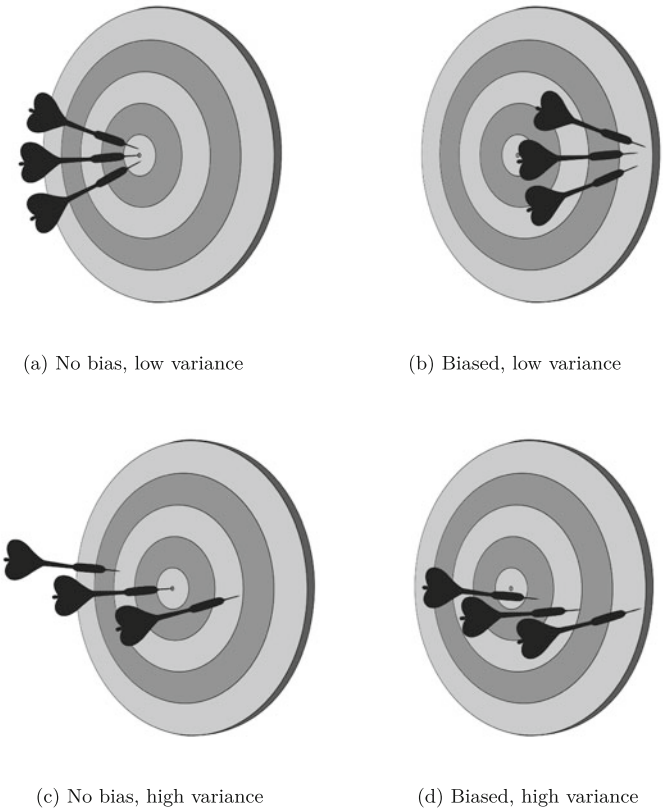
and

$$\text{Var}_\theta(T_1(X)) < \text{Var}_\theta(T_2(X))$$

for at least one $\theta \in \Theta$. It turns out that restricting our search of best estimators to unbiased estimators is sometimes a successful strategy; i.e. for many problems, a best or most efficient estimate can be found. If such an estimator exists, it is said to be UMVU (uniformly minimum variance unbiased). Uniformly means that it has the lowest variance among all other unbiased estimators for estimating the population parameter(s) $\theta$.

Consider the illustration in Fig. 9.1 to better understand the introduced concepts. Suppose we throw three darts at a target and the goal is to hit the centre of the target, i.e. the innermost circle of the dart board. The centre represents the population parameter $\theta$. The three darts play the role of three estimates $\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3$ (based on different realizations of the sample) of the population parameter $\theta$. Four possible situations are illustrated in Fig. 9.1. For example, in Fig. 9.1b, we illustrate the case of an estimator which is biased but has low variance: all three darts are "far" away from the centre of the target, but they are "close" together. If we look at Fig. 9.1a, c, we see that all three darts are symmetrically grouped around the centre of the target, meaning that there is no bias; however, in Fig. 9.1a there is much higher precision than in Fig. 9.1c. It is obvious that Fig. 9.1a presents an ideal situation: an estimator which is unbiased and has minimum variance.

**Theorem 9.2.1** *Let $X = (X_1, X_2 \ldots, X_n)$ be an i.i.d. (random) sample of a random variable $X$ with population mean $E(X_i) = \mu$ and population variance $Var(X_i) = \sigma^2$, for all $i = 1, 2, \ldots, n$. Then the arithmetic mean $\bar{X} = \sum_{i=1}^{n} X_i$ is an unbiased estimator of $\mu$ and the sample variance $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$ is an unbiased estimator of $\sigma^2$.*



(a) No bias, low variance

(b) Biased, low variance

(c) No bias, high variance

(d) Biased, high variance

**Fig. 9.1** Illustration of bias and variance

Note that the theorem holds, in general, for i.i.d. samples, irrespective of the choice of the distribution of the $X_i$'s. Note again that we are looking at the situation *before* we have any observations on $X$. Therefore, we again use capital letters to denote that the $X_i$'s are random variables which are not known beforehand (i.e. before we actually record the observations on our selected sampling units).

*Remark 9.2.1* The empirical variance $\tilde{S}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is a biased estimate of $\sigma^2$ and its bias is $-\frac{1}{n}\sigma^2$.

*Example 9.2.1* Let $X_1, X_2 \ldots, X_n$ be identically and independently distributed variables whose population mean is $\mu$ and population variance is $\sigma^2$. Then $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is an unbiased estimator of $\mu$. This can be shown as follows:

$$\mathrm{E}(\bar{X}) = \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \overset{(7.29)}{=} \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(X_i) = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu.$$

The variance of $\bar{X}$ can be calculated as follows:

$$\begin{aligned}\mathrm{Var}(\bar{X}) &= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i)\,, \quad \left[\mathrm{Cov}(X_i, X_j) = 0 \text{ using independence of } X_i\text{'s}\right] \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{\sigma^2}{n}\,.\end{aligned}$$

We conclude that $\bar{X}$ is an unbiased estimator of $\mu$ and its variance is $\frac{\sigma^2}{n}$ irrespective of the choice of the distribution of $X$. We have learned about the distribution of $\bar{X}$ already in Chap. 8, see also Appendix C.3 for the Theorem of Large Numbers and the Central Limit Theorem; however, we would like to highlight the property of "unbiasedness" in the current context.

Now, we consider another example to illustrate that estimators may not always be unbiased but may have the same variance.

*Example 9.2.2* Let $X_1, X_2 \ldots, X_n$ be identically and independently distributed variables whose population mean is $\mu$ and population variance is $\sigma^2$. Then $\tilde{X} = \bar{X} + 1 = \frac{1}{n}\sum_{i=1}^{n}(X_i + 1)$ is a biased estimator of $\mu$. This can be shown as follows:

$$\begin{aligned}\mathrm{E}(\tilde{X}) &\overset{(7.31)}{=} \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right) + \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n}1\right) \\ &\overset{(7.29)}{=} \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(X_i) + \frac{1}{n}\cdot n = \frac{1}{n}\sum_{i=1}^{n}\mu + 1 \\ &= \mu + 1 \neq \mu\,.\end{aligned}$$

However, the variance of $\tilde{X}$ is

$$\mathrm{Var}(\tilde{X}) = \mathrm{Var}\left(\bar{X} + 1\right) \overset{(7.34)}{=} \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}\,.$$

If we compare the two estimators $\tilde{X} = \frac{1}{n}\sum_{i=1}^{n}(X_i + 1)$ and $\bar{X} = \frac{1}{n}\sum_{i=1}^{n}(X_i)$, we see that both have the same variance but the former $(\tilde{X})$ is biased. The efficiency of both estimators is thus the same. It further follows that the mean squared error of $\bar{X}$ is smaller than the mean squared error of $\tilde{X}$ because the MSE consists of the sum of the variance and the squared bias. Therefore $\bar{X}$ is MSE-better than $\tilde{X}$. The comparison of bias, variance and MSE tells us that we should prefer $\bar{X}$ over $\tilde{X}$ when estimating the population mean. This is intuitive, but the argument we make is a purely statistical one.

Theorem 9.2.1 contains the following special cases:

- The sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n}X_i$ based on an i.i.d. random sample $X_1, X_2, \ldots, X_n$ from a normally distributed population $N(\mu, \sigma^2)$ is an unbiased point estimator of $\mu$.

- The sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ based on an i.i.d. random sample $X_1, X_2, \ldots, X_n$ from a normally distributed population $N(\mu, \sigma^2)$ is an unbiased point estimator of $\sigma^2$. The sample variance $\tilde{S}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is a biased estimator for $\sigma^2$, but it is asymptotically unbiased in the sense that its bias tends to zero as the sample size $n$ tends to infinity.
- The sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ based on an i.i.d. random sample $X_1, X_2, \ldots, X_n$ from a Bernoulli distributed population $B(1, p)$ is an unbiased point estimator of the probability $p$.

For illustration, we show the validity of the third statement. Let us consider an i.i.d. random sample $X_i$, $i = 1, 2 \ldots, n$, from a Bernoulli distribution, where $X_i = 1$ if an event occurs and $X_i = 0$ otherwise. Here, $p$ is the probability of occurrence of an event in the population, i.e. $p = P(X_i = 1)$. Note that $p$ is also the population mean: $\mathrm{E}(X_i) = 1 \cdot p + 0 \cdot (1 - p) = p, i = 1, 2, \ldots, n$. The arithmetic mean (relative frequency) is an unbiased estimator of $p$ because

$$\mathrm{E}(\bar{X}) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(X_i) = \frac{1}{n}\sum_{i=1}^{n} p = p,$$

and thus, we can write the estimate of $p$ as

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{9.6}$$

*Example 9.2.3* Suppose a random sample of size $n = 20$ of the weight of 10-year-old children in a particular city is drawn. Let us assume that the children's weight in the population follows a normal distribution $N(\mu, \sigma^2)$. The sample provides the following values of weights (in kg):

40.2, 32.8, 38.2, 43.5, 47.6, 36.6, 38.4, 45.5, 44.4, 40.3
34.6, 55.6, 50.9, 38.9, 37.8, 46.8, 43.6, 39.5, 49.9, 34.2

To obtain an estimate of the population mean $\mu$, we calculate the arithmetic mean of the observations as

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{20}(40.2 + 32.8 + \cdots + 34.2) = 41.97,$$

because it is an unbiased estimator of $\mu$. Similarly, we use $S^2$ to estimate $\sigma^2$ because it is unbiased in comparison to $\tilde{S}^2$. Using $s_X^2$ as an estimate for $\sigma^2$ for the given observations, we get

$$\begin{aligned}\hat{\sigma}^2 = s_x^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\ &= \frac{1}{19}((40.2 - 41.97)^2 + \cdots + (34.2 - 41.97)^2) \approx 36.85.\end{aligned}$$

The square root of 36.85 is approximately 6.07 which is the standard deviation. Note that the standard deviation based on the sample values divided by the square root of the sample size, i.e. $\hat{\sigma}/\sqrt{20}$, is called the **standard error** of the mean $\bar{X}$ (SEM). As already introduced in Chap. 3, we obtain these results in *R* using the `mean` and `var` commands.

*Example 9.2.4* A library draws a random sample of size $n = 100$ members from the members' database to see how many members have to pay a penalty for returning books late, i.e. $x_i = 1$. It turns out that 39 members in the sample have to pay a penalty. Therefore, an unbiased estimator of the population proportion of all members of the library who return books late is

$$\hat{p} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{100} \cdot 39 = \frac{39}{100} = 0.39.$$

*Remark 9.2.2* Unbiasedness and efficiency can also be defined asymptotically: we say, for example, that an estimator is asymptotically unbiased, if the bias approaches zero when the sample size tends to infinity. The concept of asymptotic efficiency involves some mathematical knowledge which is beyond the intended scope of this book. Loosely speaking, an asymptotic efficient estimator is an estimator which achieves the lowest possible (asymptotic) variance under given distributional assumptions. The estimators introduced in Sect. 9.3.1, which are based on the maximum likelihood principle, have these properties (under certain mathematically defined regularity conditions).

Next, we illustrate the properties of consistency and sufficiency of an estimator.

### 9.2.2 Consistency of Estimators

For a good estimator, as the sample size increases, the values of the estimator should get closer to the parameter being estimated. This property of estimators is referred to as consistency.

**Definition 9.2.6** Let $T_1, T_2, \ldots, T_n$, be a sequence of estimators for the parameter $\theta$ where $T_n = T_n(X_1, X_2, \ldots, X_n)$ is a function of $X_1, X_2, \ldots, X_n$. The sequence $\{T_n\}$ is a **consistent** sequence of estimators for $\theta$ if for every $\epsilon > 0$,

$$\lim_{n\to\infty} P\left[|T_n - \theta| < \epsilon\right] = 1$$

or equivalently

$$\lim_{n\to\infty} P\left[|T_n - \theta| \geq \epsilon\right] = 0.$$

This definition says that as the sample size $n$ increases, the probability that $T_n$ is getting closer to $\theta$ is approaching 1. This means that the estimator $T_n$ is getting closer to the parameter $\theta$ as $n$ grows larger. Note that there is no information on how *fast* $T_n$ is converging to $\theta$ in the sense of convergence defined above.

*Example 9.2.5* Let $X_1, X_2 \ldots, X_n$ be identically and independently distributed variables with expectation $\mu$ and variance $\sigma^2$. Then for $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, we have $\mathrm{E}(\bar{X}_n) = \mu$ and $\mathrm{Var}(\bar{X}_n) = \sigma^2/n$. For any $\epsilon > 0$, we can write the following:

$$P\left[|\bar{X}_n - \mu| \geq \epsilon\right] = P\left[|\bar{X}_n - \mu| \geq \frac{c\sigma}{\sqrt{n}}\right]$$

where $\epsilon = c\sigma/\sqrt{n}$. Using Tschebyschev's inequality (Theorem 7.4.1, p. 139), we get $\frac{1}{c^2} = \sigma^2/n\epsilon^2$, and therefore

$$P\left[|\bar{X}_n - \mu| \geq \frac{c\sigma}{\sqrt{n}}\right] \leq \frac{1}{c^2} = \frac{\sigma^2}{n\epsilon^2}$$

and

$$\lim_{n\to\infty} P\left[|\bar{X}_n - \mu| \geq \frac{c\sigma}{\sqrt{n}}\right] \leq \lim_{n\to\infty} \frac{\sigma^2}{n\epsilon^2} = 0,$$

provided $\sigma^2$ is finite. Hence $\bar{X}_n, n = 1, 2, \ldots,$ converges to $\mu$ and therefore $\bar{X}_n$ is a consistent estimator of $\mu$.

*Remark 9.2.3* We call this type of consistency *weak* consistency. Another definition is *MSE* consistency, which says that an estimator is MSE consistent if $MSE \longrightarrow 0$ as $n \to \infty$. If the estimator is unbiased, it is sufficient that $\mathrm{Var} \longrightarrow 0$ as $n \to \infty$. If $T_n(X)$ is MSE consistent, it is also weakly consistent. Therefore, it follows that an unbiased estimator with its variance approaching zero as the sample size approaches infinity is both MSE consistent and weakly consistent.

In Example 9.2.5, the variance of $T_n(X) = \bar{X}_n$ is $\sigma^2/n$ which goes to zero as $n$ goes to $\infty$ and therefore $\bar{X}_n$ is both weakly consistent and MSE consistent.

### 9.2.3 Sufficiency of Estimators

Sufficiency is another criterion to judge the quality of an estimator. Before delving deeper into the subject matter, we first try to understand some basic concepts.

Consider two independent random variables $X$ and $Y$, each following a $N(\mu, 1)$ distribution. We conclude that both $X$ and $Y$ contain information about $\mu$. Consider two estimators of $\mu$ as $\hat{\mu}_1 = X + Y$ and $\hat{\mu}_2 = X - Y$. Suppose we want to know whether to use $\hat{\mu}_1$ or $\hat{\mu}_2$ to estimate $\mu$. We notice that $\mathrm{E}(\hat{\mu}_1) = \mathrm{E}(X) + \mathrm{E}(Y) = \mu + \mu = 2\mu$, $\mathrm{E}(\hat{\mu}_2) = \mathrm{E}(X) - \mathrm{E}(Y) = \mu - \mu = 0$, $\mathrm{Var}(\hat{\mu}_1) = \mathrm{Var}(X) + \mathrm{Var}(Y) = 1 + 1 = 2$ and $\mathrm{Var}(\hat{\mu}_2) = \mathrm{Var}(X) + \mathrm{Var}(Y) = 1 + 1 = 2$. Using the additivity property of the normal distribution, which was introduced in Remark 8.2.2, we can say that $\hat{\mu}_1 \sim N(2\mu, 2)$ and $\hat{\mu}_2 \sim N(0, 2)$. So $\hat{\mu}_1$ contains information about $\mu$, whereas $\hat{\mu}_2$

does not contain any information about $\mu$. In other words, $\hat{\mu}_2$ loses the information about $\mu$. We call this property "loss of information".

If we want to make conclusions about $\mu$ using both $X$ and $Y$, we need to acknowledge that the dimension of them is 2. On the other hand, if we use $\hat{\mu}_1$ or equivalently $\hat{\mu}_1/2 \sim N(\mu, \frac{1}{2})$, then we need to concentrate only on one variable and we say that it has dimension 1. It follows that $\hat{\mu}_1$ and $\hat{\mu}_1/2$ provide the same information about $\mu$ as provided by the entire sample on both $X$ and $Y$. So we can say that either $\hat{\mu}_1$ or $\hat{\mu}_1/2$ is sufficient to provide the same information about $\mu$ that can be obtained on the basis of the entire sample. This is the idea behind the concept of sufficiency and it results in the reduction of dimension. In general, we can say that if all the information about $\mu$ contained in the sample of size $n$ can be obtained, for example, through the sample mean then it is sufficient to use this one-dimensional summary statistic to make inference about $\mu$.

**Definition 9.2.7** Let $X_1, X_2, \ldots, X_n$ be a random sample from a probability density function (or probability mass function) $f(x, \theta)$. A statistic $T$ is said to be sufficient for $\theta$ if the conditional distribution of $X_1, X_2, \ldots, X_n$ given $T = t$ is independent of $\theta$.

The Neyman–Fisher Factorization Theorem provides a practical way to find sufficient statistics.

**Theorem 9.2.2** (Neyman–Fisher Factorization Theorem (NFFT)) *Let $X_1, X_2, \ldots, X_n$ be a random sample from a probability density function (or probability mass function) $f(x, \theta)$. A statistic $T = T(x_1, x_2, \ldots, x_n)$ is said to be sufficient for $\theta$ if and only if the joint density of $X_1, X_2, \ldots, X_n$ can be factorized as*

$$f(x_1, x_2, \ldots, x_n; \theta) = g(t, \theta) \cdot h(x_1, x_2, \ldots, x_n)$$

*where $h(x_1, x_2, \ldots, x_n)$ is nonnegative and does not involve $\theta$; and $g(t, \theta)$ is a nonnegative function of $\theta$ which depends on $x_1, x_2, \ldots, x_n$ only through $t$, which is a particular value of $T$.*

*This theorem holds for discrete random variables too. Any one-to-one function of a sufficient statistic is also sufficient. A function $f$ is called one-to-one if whenever $f(a) = f(b)$ then $a = b$.*

*Example 9.2.6* Let $X_1, X_2, \ldots, X_n$ be a random sample from $N(\mu, 1)$ where $\mu$ is unknown. We attempt to find a sufficient statistic for $\mu$. Consider the following function as the joint distribution of $x_1, x_2, \ldots, x_n$ (whose interpretation will become clearer in the next section):

$$\begin{aligned} f(x_1, x_2, \ldots, x_n; \mu) &= \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right) \\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\frac{n\mu^2}{2} + \mu\sum_{i=1}^{n} x_i\right)\exp\left(-\frac{1}{2}\sum_{i=1}^{n} x_i^2\right). \end{aligned}$$

Here

$$g(t, \mu) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\frac{n\mu^2}{2} + \mu \sum_{i=1}^{n} x_i\right),$$

$$h(x_1, x_2, \ldots, x_n) = \exp\left(-\frac{1}{2}\sum_{i=1}^{n} x_i^2\right),$$

$$t = t(x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} x_i.$$

Using the Neyman–Fisher Factorization Theorem, we conclude that $T = T(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} X_i$ is a sufficient statistic for $\mu$. Also, $T = T(X_1, X_2, \ldots, X_n) = \bar{X}$ is sufficient for $\mu$ as it is a one-to-one statistic of $\sum_{i=1}^{n} X_i$. On the other hand, $T = \bar{X}^2$ is not sufficient for $\mu$ as it is not a one-to-one function of $\sum_{i=1}^{n} X_i$. The important point here is that $\bar{X}$ is a function of the sufficient statistic and hence a good estimator for $\mu$. It is thus summarizing the sample information about the parameter of interest in a complete yet parsimonious way. Another, multivariate, example of sufficiency is given in Appendix C.4.

## 9.3 Point Estimation

In the previous section, we introduced and discussed various properties of estimators. In this section, we want to show how one can find estimators with good properties. In the general case, properties such as unbiasedness and efficiency cannot be guaranteed for a finite sample. But often, the properties can be shown to hold asymptotically.

### 9.3.1 Maximum Likelihood Estimation

We have used several estimators throughout the book without stating explicitly that they are estimators. For example, we used the sample mean ($\bar{X}$) to estimate $\mu$ in a $N(\mu, \sigma^2)$ distribution; we also used the sample proportion (relative frequency) to estimate $p$ in a $B(1, p)$ distribution, etc. The obvious question is how to obtain a good statistic to estimate an unknown parameter, for example how to determine that the sample mean can be used to estimate $\mu$. We need a general framework for parameter estimation. The method of **maximum likelihood** provides such an approach. For the purpose of illustration, we introduce the method of maximum likelihood estimation with an example using the Bernoulli distribution.

*Example 9.3.1* Consider an i.i.d. random sample $X = (X_1, X_2, \ldots, X_n)$ from a Bernoulli population with $p = P(X_i = 1)$ and $(1 - p) = P(X_i = 0)$. The joint probability mass function for a given set of realizations $x_1, x_2, \ldots, x_n$ (i.e. the data) is

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n|p) = P(X_1 = x_1|p) \cdot \cdots \cdot P(X_n = x_n|p)$$
$$= \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}. \tag{9.7}$$

This is a function of $(x_1, x_2, \ldots, x_n)$ given the parameter $p$. The product results from the fact that the draws are independent and the fact that $p^{x_i}(1-p)^{1-x_i} = p$ if $x_i = 1$ and $p^{x_i}(1-p)^{1-x_i} = 1-p$ if $x_i = 0$. That is, the term $p^{x_i}(1-p)^{1-x_i}$ covers results from both possible outcomes. Now, consider a random sample where the values $x = (x_1, x_2, \ldots, x_n)$ are known, for example $x = (0, 1, 0, 0, \ldots, 1)$. Then, (9.7) can be seen as a function of $p$ because $(x_1, x_2, \ldots, x_n)$ is known. In this case, after obtaining a sample of data, the function is called the likelihood function and can be written as

$$L(x_1, x_2, \ldots, x_n|p) = \prod_{i=1}^{n} p^{x_1}(1-p)^{1-x_i}\ . \tag{9.8}$$

The joint density function of $X_1, X_2, \ldots, X_n$ is called the **likelihood function**. For better understanding, consider a sample of size 5 with $x = (x_1 = 1, x_2 = 1, x_3 = 0, x_4 = 1, x_5 = 0)$. The likelihood (function) is

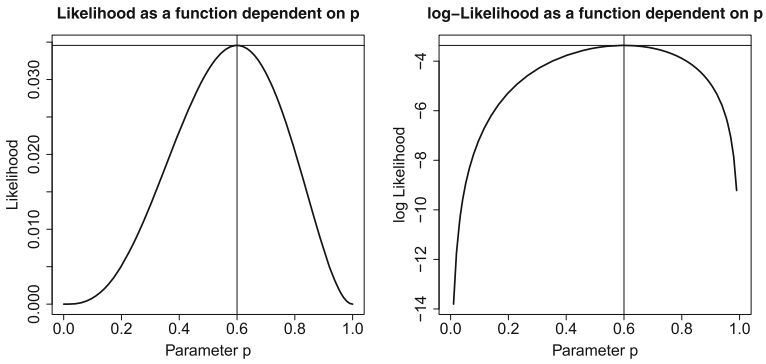
$$L(1, 1, 0, 1, 0|p) = p \cdot p \cdot (1-p) \cdot p \cdot (1-p) = p^3(1-p)^2\ . \tag{9.9}$$

The maximum likelihood estimation principle now says that the estimator $\hat{p}$ of $p$ is the value of $p$ which maximizes the likelihood (9.8) or (9.9). In other words, the maximum likelihood estimate is the value which maximizes the probability of observing the realized sample from the likelihood function. In general, i.e. for any sample, we have to maximize the likelihood function (9.9) with respect to $p$. We use the well-known principle of maxima–minima to maximize the likelihood function in this case. In principle, any other optimization procedure can also be used, for example numerical algorithms such as the Newton–Raphson algorithm. If the likelihood is differentiable, the first-order condition for the maximum is that the first derivative with respect to $p$ is zero. For maximization, we can transform the likelihood by a strictly monotone increasing function. This guarantees that the potential maximum is taken at the same point as in the original likelihood. A good and highly common choice is the *natural logarithm* since it transforms products in sums and sums are easy to differentiate by differentiating each term in the sum. The log-likelihood in our example is therefore

$$l(1, 1, 0, 1, 0|p) = \ln L(1, 1, 0, 1, 0, |p) = \ln\left\{p^3(1-p)^2\right\} \tag{9.10}$$
$$= 3\ln(p) + 2\ln(1-p) \tag{9.11}$$

where ln denotes the natural logarithm function and we use the rules

$$\ln(a \cdot b) = \ln(a) + \ln(b)\ , a > 0, b > 0$$
$$\ln\left(\frac{a}{b}\right) = \ln(a) - \ln(b)\ , a > 0, b > 0$$
$$\ln\left(a^b\right) = b\ln(a)\ , a > 0\ .$$




(a) Likelihood (b) Log-likelihood

**Fig. 9.2** Illustration of the likelihood and log-likelihood function of a binomial distribution

Taking the first derivative of (9.10) with respect to $p$ results in

$$\frac{\partial l(1,1,0,1,0|p)}{\partial p} = \frac{3}{p} - \frac{2}{1-p} \,. \tag{9.12}$$

Setting (9.12) to zero and solving for $p$ leads to

$$\begin{aligned} \frac{3}{p} - \frac{2}{1-p} &= 0 \\ \frac{3}{p} &= \frac{2}{1-p} \\ 3(1-p) &= 2p \\ 5p &= 3 \\ \hat{p}_{\text{ML}} = \frac{3}{5} = \frac{1}{5}(1+1+0+1+0) &= \bar{x} \,. \end{aligned}$$

The value of the second-order partial derivative of (9.9) with respect to $p$ at $p = \hat{p}_{\text{ML}}$ is negative which ensures that $\hat{p}_{\text{ML}}$ maximizes the likelihood function. It follows from this example that the maximum likelihood estimate for $p$ leads to the well-known arithmetic mean. Figure 9.2 shows the likelihood function and the log-likelihood function as functions of $p$, where $p \in [0, 1]$. The figures show that the likelihood function and the log-likelihood function have the same maxima at $p = 3/5 = 0.6$.

Maximum likelihood estimators have some important properties: they are usually consistent, asymptotically unbiased, asymptotically normally distributed, asymptotically efficient, and sufficient. Even if they are not, a function of a sufficient statistic can always be found which has such properties. This is the reason why maximum likelihood estimation is popular. By "asymptotically" we mean that the properties hold as $n$ tends to infinity, i.e. as the sample size increases. There might be other good estimators in a particular context, for example estimators that are efficient and not only asymptotically efficient; however, in general, the ML principle is a great

choice in many circumstances. We are going to use it in the following sections and chapters, for instance for general point estimation and in the linear regression model (Chap. 11).

*Remark 9.3.1* More examples of maximum likelihood estimators are given in Exercises 9.1–9.3.

### 9.3.2 Method of Moments

The **method of moments** is another well-known method to derive the estimators for population parameters. Below, we outline this principle briefly by way of example.

The idea is that the population parameters of interest can be related to the moments (e.g. expectation, variance) of the distribution of the considered random variables.

A simple case is the estimator for the expected value $\mathrm{E}(X) = \mu$ of a population using an i.i.d. random sample $X = (X_1, \ldots, X_n)$. In this case, $\hat{\mu} = \bar{X}$ is the natural moment estimator of $\mu$. Further, since $\mathrm{E}(X^2) = \sigma^2 + \mu^2$, an estimator of $\sigma^2 + \mu^2$ is $\frac{1}{n}\sum_{i=1}^n X_i^2$. Using $\bar{X}^2$ as an estimator for $\mu^2$, this results in the biased, but asymptotically unbiased estimate

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \left(\frac{1}{n}\sum_{i=1}^n X_i\right)^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2 \,.$$

An extension of this method is the *generalized method of moments* (GMM). GMM estimators have interesting properties: under relatively week conditions (not further discussed here), they are consistent and asymptotically normal, as well as efficient in the class of those estimators that do not use any additional information besides the information included in the moment conditions. Usually, they require a two-step estimating approach or an iterative estimating procedure.

The **least squares estimator** for a linear regression model with i.i.d. random errors, discussed in detail in Chap. 11, can be seen as a special case of a GMM estimator.

## 9.4 Interval Estimation

### 9.4.1 Introduction

Let us first consider an example to understand what we mean by interval estimation. Consider a situation in which a lady wants to know the time taken to travel from her home to the train station. Suppose she makes 20 trips and notes down the time taken. To get an estimate of the expected time, one can use the arithmetic mean. Let us say $\bar{x} = 25$ min. This is the point estimate for the expected travelling time. It may not be appropriate to say that she will always take exactly 25 min to reach the train station.

Rather the time may vary by a few minutes each time. To take this into account, the time can be estimated in the form of an interval: it may then be found that the time varies mostly between 20 and 30 min. Such a statement is more informative. Both expectation and variation of the data are taken into account. The interval (20, 30 min) provides a range in which most of the values are expected to lie. We call this concept interval estimation.

A point estimate on its own does not take into account the precision of the estimate. The deviation between the point estimate and the true parameter (e.g. $|\bar{x} - \mu|$) can be substantial, especially when the sample size is small. To incorporate the information about the precision of an estimate in the estimated value, a **confidence interval** can be constructed. It is a **random interval** with **lower and upper bounds**, $I_l(\mathbf{X})$ and $I_u(\mathbf{X})$, such that the unknown parameter $\theta$ is covered by a prespecified probability of at least $1 - \alpha$:

$$P_\theta(I_l(\mathbf{X}) \leq \theta \leq I_u(\mathbf{X})) \geq 1 - \alpha. \tag{9.13}$$

The probability $1 - \alpha$ is called the **confidence level** or **confidence coefficient**, $I_l(\mathbf{X})$ is called the **lower confidence bound** or **lower confidence limit** and $I_u(\mathbf{X})$ is called the **upper confidence bound** or **upper confidence limit**. It is important to note that the bounds are random and the parameter is a fixed value. This is the reason why we say that the true parameter is covered by the interval with probability $1 - \alpha$ and **not** that the probability that the interval contains the parameter is $1 - \alpha$. Please note that some software packages use the term "error bar" when referring to confidence intervals.

*Frequency interpretation of the confidence interval:* Suppose $N$ independent samples $\mathbf{X}^{(j)}$, $j = 1, 2, \ldots, N$, of size $n$ are sampled from the same population and $N$ confidence intervals of the form $[I_l(\mathbf{X}^{(j)}), I_u(\mathbf{X}^{(j)})]$ are calculated. If $N$ is large enough, then on an average $N(1 - \alpha)$ of the intervals (9.13) cover the true parameter.

*Example 9.4.1* Let a random variable follow a normal distribution with $\mu = 10$ and $\sigma^2 = 1$. Suppose we draw a sample of $n = 10$ observations repeatedly. The sample will differ in each draw, and hence, the mean and the confidence interval will also differ. The data sets are realizations from random variables. Have a look at Fig. 9.3 which illustrates the mean and the 95 % confidence intervals for 6 random samples. They vary with respect to the mean and the confidence interval width. Most of the means are close to $\mu = 10$, but not all. Similarly, most confidence intervals, but not all, include $\mu$. This is the idea of the frequency interpretation of the confidence interval: different samples will yield different point and interval estimates. Most of the times the interval will cover $\mu$, but not always. The coverage probability is specified by $1 - \alpha$, and the frequency interpretation means that we expect that (approximately) $(1 - \alpha) \cdot 100\,\%$ of the intervals to cover the true parameter $\mu$. In that sense, the location of the interval will give us some idea about where the true but unknown population parameter $\mu$ lies, while the length of the interval reflects our uncertainty about $\mu$: the wider the interval is, the higher is our uncertainty about the location of $\mu$.

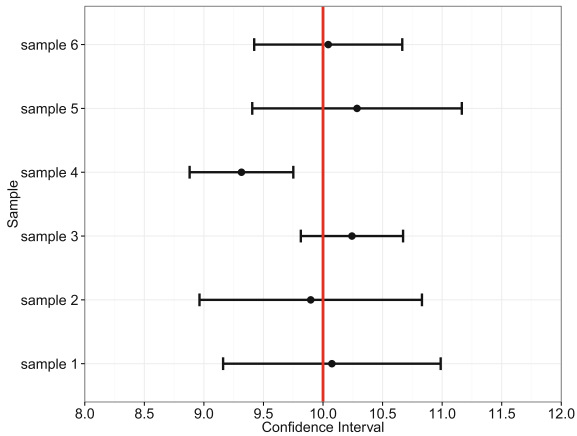



**Fig. 9.3** Frequency interpretation of confidence intervals

We now introduce the following confidence intervals:

- Confidence interval for the mean $\mu$ of a normal distribution.
- Confidence interval for the probability $p$ of a binomial random variable.
- Confidence interval for the odds ratio.

### 9.4.2 Confidence Interval for the Mean of a Normal Distribution

**Confidence Interval for $\mu$ When $\sigma^2 = \sigma_0^2$ is Known.**

Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from a $N(\mu, \sigma_0^2)$ distribution where $\sigma_0^2$ is assumed to be known. We use the point estimate $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ to estimate $\mu$ and construct a confidence interval around the mean $\mu$. Using the Central Limit Theorem (Appendix C.3, p. 426), it follows that $\bar{X}$ follows a $N(\mu, \sigma_0^2/n)$ distribution. Therefore $\sqrt{n}(\bar{X} - \mu)/\sigma_0 \sim N(0, 1)$, and it follows that

$$P_\mu\left(\left|\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma_0}\right| \leq z_{1-\frac{\alpha}{2}}\right) = 1 - \alpha \tag{9.14}$$

where $z_{1-\alpha/2}$ denotes the $(1 - \alpha/2)$ quantile of the standard normal distribution $N(0, 1)$. We solve this inequality for the unknown $\mu$ and get the desired confidence interval as follows:

$$P_\mu\left[-z_{1-\frac{\alpha}{2}} \leq \left(\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma_0}\right) \leq z_{1-\frac{\alpha}{2}}\right] = 1 - \alpha$$

or

$$P_\mu\left[\bar{X} - z_{1-\alpha/2}\frac{\sigma_0}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{1-\alpha/2}\frac{\sigma_0}{\sqrt{n}}\right] = 1 - \alpha\,.$$

The confidence interval for $\mu$ is thus obtained as

$$[I_l(\mathbf{X}), I_u(\mathbf{X})] = \left[\bar{X} - z_{1-\alpha/2}\frac{\sigma_0}{\sqrt{n}}, \bar{X} + z_{1-\alpha/2}\frac{\sigma_0}{\sqrt{n}}\right]. \tag{9.15}$$

This is known as $(1-\alpha)\%$ confidence interval for $\mu$ or the confidence interval for $\mu$ with confidence coefficient $\alpha$.

We can use the *R* function `qnorm` or Table C.1 to obtain $z_{1-\frac{\alpha}{2}}$, see also Sects. 8.4, A.3, and C.7. For example, for $\alpha = 0.05$ and $\alpha = 0.01$ we get $z_{1-\frac{\alpha}{2}} = z_{0.975} = 1.96$ and $z_{1-\frac{\alpha}{2}} = z_{0.995} = 2.576$ using `qnorm(0.975)` and `qnorm(0.995)`. This gives us the quantiles we need to determine a 95 % and 99 % confidence interval, respectively.

*Example 9.4.2* We consider again Example 9.2.3 where we evaluated the weight of 10-year-old children. Assume that the variance is known to be 36; then the upper and lower limits of a 95 % confidence interval for the expected weight $\mu$ can be calculated as follows:

$$I_l(X) = \bar{X} - z_{1-\alpha/2}\frac{\sigma_0}{\sqrt{n}} = 41.97 - 1.96\frac{\sqrt{36}}{\sqrt{20}} \approx 39.34\,,$$
$$I_u(X) = \bar{X} + z_{1-\alpha/2}\frac{\sigma_0}{\sqrt{n}} = 41.97 + 1.96\frac{\sqrt{36}}{\sqrt{20}} \approx 44.59.$$

We get the confidence interval $[I_u(X), I_o(X)] = [39.34, 44.59]$. With 95 % confidence, the true parameter $\mu$ is covered by the interval [39.34, 44.59].

**Confidence Interval for $\mu$ When $\sigma^2$ is Unknown.**

Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from $N(\mu, \sigma^2)$ where $\sigma^2$ is assumed to be unknown and is being estimated by the sample variance $S_X^2$. We know from Sect. 8.3.1 that

$$\frac{(n-1)S_X^2}{\sigma^2} \sim \chi^2_{n-1}\,.$$

It can be shown that $\bar{X}$ and $S_X^2$ are stochastically independent. Thus, we know that

$$\frac{\sqrt{n}(\bar{X}-\mu)}{S_X} \sim t_{n-1}$$

follows a $t$-distribution with $n-1$ degrees of freedom. We can use this result to determine the confidence interval for $\mu$ as

$$P_\mu\left[-t_{1-\frac{\alpha}{2},n-1} \le \left(\frac{\sqrt{n}(\bar{X}-\mu)}{S_X}\right) \le t_{1-\frac{\alpha}{2},n-1}\right] = 1-\alpha$$

or

$$P_\mu\left[\bar{X} - t_{1-\frac{\alpha}{2},n-1}\frac{S_X}{\sqrt{n}} \le \mu \le \bar{X} + t_{1-\frac{\alpha}{2},n-1}\frac{S_X}{\sqrt{n}}\right] = 1-\alpha\,.$$

The confidence interval for $\mu$ is thus obtained as

$$[I_l(\mathbf{X}),\, I_u(\mathbf{X})] = \left[\bar{X} - t_{n-1;1-\alpha/2} \cdot \frac{S_X}{\sqrt{n}},\ \bar{X} + t_{n-1;1-\alpha/2} \cdot \frac{S_X}{\sqrt{n}}\right] \tag{9.16}$$

which is the $100(1-\alpha)\%$ confidence interval for $\mu$ or the confidence interval for $\mu$ with confidence coefficient $\alpha$.

The interval (9.16) is, in general, wider than the interval (9.15) for identical $\alpha$ and identical sample size $n$, since the unknown parameter $\sigma^2$ is estimated by $S_X^2$ which induces additional uncertainty. The quantiles for the $t$-distribution can be obtained using the *R* command `qt` or Table C.2.

*Example 9.4.3* Consider Example 9.4.2 where we evaluated the weight of 10-year-old children. We have already calculated the point estimate of $\mu$ as $\bar{x} = 41.97$. With $t_{19;0.975} = 2.093$, obtained via `qt(0.975,19)` or Table C.2, the upper and lower limits of a 95 % confidence interval for $\mu$ are obtained as

$$I_u(X) = \bar{x} - t_{19;0.975} \cdot \frac{s_X}{\sqrt{n}} = 41.97 - 2.093 \cdot \frac{6.07}{\sqrt{20}} \approx 39.12\ ,$$
$$I_o(X) = \bar{x} + t_{19;0.975} \cdot \frac{s_X}{\sqrt{n}} = 41.97 + 2.093 \cdot \frac{6.07}{\sqrt{20}} \approx 44.81\ .$$

Therefore, the confidence interval is $[I_l(X),\, I_u(X)] = [39.13, 44.81]$. In *R*, we can use the `conf.int` value of the `t.test` command to get a confidence interval for the mean (see also Example 10.3.3 for more details on `t.test`). The default is a 95 % confidence interval, but it can be changed easily if desired:

```
x <- c(40.2, 32.8, 38.2, 43.5, ..., 49.9, 34.2)
t.test(x,conf.level = 0.95)$conf.int
[1] 39.12384 44.80616
```

There is no unique best way to draw the calculated confidence intervals in *R*. Among many other options, one can simply work with the `plot` functionality or use `geom_errorbar` in conjunction with a `ggplot` object created with the library `ggplot2`, or use the `plotCI` command in the library `plotrix`.

### 9.4.3 Confidence Interval for a Binomial Probability

Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample from a Bernoulli distribution $B(1, p)$. Then $Y = \sum_{i=1}^{n} X_i$ has a binomial distribution $B(n, p)$.

We have already introduced $\hat{p}$ as an estimator for $p$:

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n} Y.$$

From (8.8), we know that $\text{Var}(Y) = np(1-p)$. Applying rule (7.33), the variance of the estimator $\hat{p}$ is

$$\text{Var}(\hat{p}) = \frac{p(1-p)}{n}$$

and it can be estimated by

$$S_{\hat{p}}^2 = \frac{\hat{p}(1-\hat{p})}{n}.$$

Nowadays, the exact confidence intervals of the binomial distribution function can be easily calculated using computer implementations. Nevertheless, (i) for a sufficiently large sample size $n$, (ii) if $p$ is not extremely low or high, and (iii) if the condition $np(1-p) \geq 9$ is fulfilled, we can use an approximation based on the normal distribution to calculate confidence intervals. To be more specific, one can show that

$$Z = \frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})/n}} \overset{approx.}{\sim} N(0,1). \tag{9.17}$$

This gives us

$$P\left[\hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right] \approx 1-\alpha, \tag{9.18}$$

and we get a confidence interval for $p$ as

$$\left[\hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]. \tag{9.19}$$

*Example 9.4.4* We look again at Example 9.2.4 where we evaluated the proportion of members who had to pay a penalty. Out of all borrowers, 39 % brought back their books late and thus had to pay a fee. A 95 % confidence interval for the probability $p$ of bringing back a book late can be constructed using the normal approximation, since $n\hat{p}(1-\hat{p}) = 100 \cdot 0.39 \cdot 0.61 = 23.79 > 9$. With $z_{1-\alpha/2} = z_{0.975} = 1.96$ and $\hat{p} = 0.39$, we get the 95 % confidence interval as

$$\left[0.39 - 1.96\sqrt{\frac{0.39 \cdot 0.61}{100}},\ 0.39 + 1.96\sqrt{\frac{0.39 \cdot 0.61}{100}}\right] = [0.294, 0.486].$$

In *R*, an exact confidence interval can be found using the function `binom.test`:

```
binom.test(x=39,n=100)$conf.int
[1] 0.2940104 0.4926855
```

One can see that the exact and approximate confidence limits differ slightly due to the normal approximation which approximates the exact binomial probabilities.

### 9.4.4 Confidence Interval for the Odds Ratio

In Chap. 4, we introduced the odds ratio to determine the strength of association between two binary variables. One may be interested in the dispersion of the odds ratio and hence calculate a confidence interval for it. Recall the notation for $2 \times 2$ contingency tables:

| | | $Y$ | | Total (row) |
|---|---|---|---|---|
| | | $y_1$ | $y_2$ | |
| $X$ | $x_1$ | $a$ | $b$ | $a+b$ |
| | $x_2$ | $c$ | $d$ | $c+d$ |
| Total (column) | | $a+c$ | $b+d$ | $n$ |

In the spirit of the preceding sections, we can interpret the entries in this contingency table as population parameters. For example, $a$ describes the absolute frequency of observations in the population for which $Y = y_1$ and $X = x_1$. If we have a sample then we can estimate $a$ by the number of *observed* observations $n_{11}$ for which $Y = y_1$ and $X = x_1$. We can thus view $n_{11}$ to be an estimator for $a$, $n_{12}$ to be an estimator for $b$, $n_{21}$ to be an estimator for $c$, and $n_{22}$ to be an estimator for $d$. It follows that

$$\widehat{\mathrm{OR}} = \frac{n_{11}n_{22}}{n_{12}n_{21}} \tag{9.20}$$

serves as the point estimate for the population odds ratio $\mathrm{OR} = ad/bc$. To construct a confidence interval for the odds ratio, we need to work on a log-scale. The log odds ratio,

$$\theta_0 = \ln \mathrm{OR} = \ln a - \ln b - \ln c + \ln d, \tag{9.21}$$

takes the natural logarithm of the odds ratio. It is evident that it can be estimated using the observed absolute frequencies of the joint frequency distribution of $X$ and $Y$:

$$\hat{\theta}_0 = \ln \widehat{\mathrm{OR}} = \ln \frac{n_{11}n_{22}}{n_{12}n_{21}}\,. \tag{9.22}$$

It can be shown that $\hat{\theta}_0$ follows approximately a normal distribution with expectation $\theta_0$ and standard deviation

$$\hat{\sigma}_{\hat{\theta}_0} = \left(\frac{1}{n_{11}} + \frac{1}{n_{22}} + \frac{1}{n_{12}} + \frac{1}{n_{21}}\right)^{\frac{1}{2}}\,. \tag{9.23}$$

Following the reasoning explained in the earlier section on confidence intervals for binomial probabilities, we can calculate the $100(1-\alpha)\%$ confidence interval for $\theta_0$ under a normal approximation as follows:

$$\left[\hat{\theta}_0 - z_{1-\frac{\alpha}{2}}\hat{\sigma}_{\hat{\theta}_0}, \hat{\theta}_0 + z_{1-\frac{\alpha}{2}}\hat{\sigma}_{\hat{\theta}_0}\right] = [I_u, I_o]\,. \tag{9.24}$$

Since we are interested in the confidence interval of the odds ratio, and not the log odds ratio, we need to transform back the lower and upper bound of the confidence interval as

$$\left[\exp(I_u), \exp(I_o)\right] . \tag{9.25}$$

*Example 9.4.5* Recall Example 4.2.5 from Chap. 4 where we were interested in the association of smoking with a particular disease. The data is summarized in the following $2 \times 2$ contingency table:

| | | Smoking | | Total (row) |
|---|---|---|---|---|
| | | Yes | No | |
| Disease | Yes | 34 | 66 | 100 |
| | No | 22 | 118 | 140 |
| Total (column) | | 56 | 184 | 240 |

The odds ratio was estimated to be 2.76, and we therefore concluded that the chances of having the particular disease is 2.76 times higher for smokers compared with non-smokers. To calculate a 95 % confidence intervals, we need $\hat{\theta}_0 = \ln(2.76)$, $z_{1-\frac{\alpha}{2}} \approx 1.96$ and

$$\begin{aligned}\hat{\sigma}_{\hat{\theta}_0} &= \left(\frac{1}{n_{11}} + \frac{1}{n_{22}} + \frac{1}{n_{12}} + \frac{1}{n_{21}}\right)^{\frac{1}{2}} \\ &= \left(\frac{1}{34} + \frac{1}{118} + \frac{1}{66} + \frac{1}{22}\right)^{\frac{1}{2}} \approx 0.314.\end{aligned}$$

The confidence interval for the log odds ratio is

$$[\ln(2.76) - 1.96 \cdot 0.314, \ln(2.76) + 1.96 \cdot 0.314] \approx [0.40, 1.63] .$$

Exponentiation of the confidence interval bounds yields the 95 % confidence interval for the odds ratio as

$$[1.49, 5.11] .$$

There are many ways to obtain the same results in *R*. One option is to use the `oddsratio` function of the library `epitools`. Note that we need to specify "wald" under the methods option to get confidence intervals which use the normal approximation as we did in this case.

```
library(epitools)
smd <- matrix(c(34,22,66,118),ncol=2,nrow=2)  #data
oddsratio(smd,method='wald')
```

## 9.5 Sample Size Determinations

Confidence intervals help us estimating the precision of point estimates. What if we are required to adhere to a prespecified precision level? We know that the variance decreases as the sample size increases. In turn, confidence intervals become narrower. On the other hand, increasing the sample size has its own consequences. For example, the cost and time involved in setting up experiments, or conducting a survey, increases. In these situations it is important to find a balance between the variability of the estimates and the sample size. We cannot control the variability in the data in most of the situations, but it is possible to control the sample size and therefore the precision of our estimates. For example, we can control the number of people to be interviewed in a survey—given the resources which are available. We discuss how to determine the number of observations needed to get a particular precision (length) of the confidence interval. We find the answers to such questions using the formulae for confidence intervals.

### Sample Size Calculation for $\mu$.

Let us consider the situation where we are interested in estimating the population mean $\mu$. The length of the confidence interval (9.15) for the point estimate $\bar{X}$ is

$$2z_{1-\alpha/2}\frac{\sigma_0}{\sqrt{n}}. \tag{9.26}$$

We would now like to fix the width of the confidence interval and come up with a sample size which is required to achieve this width. Let us fix the length of the confidence interval as

$$\Delta = 2z_{1-\alpha/2}\frac{\sigma_0}{\sqrt{n}}. \tag{9.27}$$

Assume we have knowledge of $\sigma_0$. The knowledge about $\sigma_0$ can be obtained, for example, through a pilot study or past experience with the experiment. We are interested in obtaining the value of $n$ for which a confidence interval has a fixed confidence width of $\Delta$ or less. Rearranging (9.27) gives us

$$n \geq \left[2\frac{z_{1-\alpha/2}\sigma_0}{\Delta}\right]^2. \tag{9.28}$$

This means a minimum or optimum sample size is

$$n_{opt} = \left[2\frac{z_{1-\alpha/2}\sigma_0}{\Delta}\right]^2. \tag{9.29}$$

The sample size $n_{opt}$ ensures that the $1-\alpha$ confidence interval for $\mu$ has at most length $\Delta$. But note that we have assumed that $\sigma_0$ is known. If we do not know $\sigma_0$ (which is more likely in practice), we have to make an assumption about it, e.g. by using an estimate from a former study, a pilot study, or other external information. Practically, (9.28) is used in the case of known and unknown $\sigma_0^2$.

*Example 9.5.1* A call centre is interested in determining the expected length of a telephone call as precisely as possible. The requirements are that the 95 % confidence interval for $\mu$ should have a width of 1 min. Suppose that the call centre has developed a pilot study in which $\sigma_0$ was estimated to be 5 min. The sample size $n$ that is needed to estimate the expected length of the phone calls with the desired precision is:

$$n \geq \left[\frac{2z_{1-\alpha/2}\sigma_0}{\Delta}\right]^2 = \left[\frac{2 \times 1.96 \times 5}{1}\right]^2 \approx 384.$$

This means that at least 384 calls are required to get the desired confidence interval width.

### Sample Size Calculation for $p$.

We can follow the earlier reasoning and determine the optimum sample size for a specific confidence interval width using the confidence interval definition (9.19). Since the width of the confidence interval is

$$\Delta = 2z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},$$

we get

$$n \geq \left[2\frac{z_{1-\alpha/2}}{\Delta}\right]^2 \hat{p}(1-\hat{p}). \tag{9.30}$$

*Example 9.5.2* A factory may be interested in the probability of an error in an operating process. The length of the confidence interval should be $\pm 2$ %, i.e. $\Delta = 0.04$. Suppose it is speculated that the error probability is 10 %; we may then use $\hat{p} = 0.1$ as our prior judgment for the true value of $p$. This yields

$$n \geq \left[2\frac{z_{1-\alpha/2}}{\Delta}\right]^2 \hat{p}(1-\hat{p}) = \left[2 \times \frac{1.96}{0.04}\right]^2 0.1 \cdot (1-0.1) \approx 865. \tag{9.31}$$

This means we need a sample size of at least 865 to obtain the desired width of the confidence interval for $p$.

The above examples for both $\mu$ and $p$ have shown us that without external knowledge about the research question of interest, it is difficult to come up with an appropriate sample size. Results may vary considerably depending on what type of information is assumed to be known. With limited knowledge, it can be useful to report results for different widths of confidence intervals and hypothesized values of $p$ or $\sigma_0$.

Sample size calculations can be highly complex in many practical situations and may not remain as simple as in the examples considered here. For example, Chap. 10 uses additional concepts in the context of hypothesis testing, such as the power, which can be taken into consideration when estimating sample sizes. However,

in this case, calculations and interpretations become more difficult and complex. A detailed overview of sample size calculations can be found in Chow et al. (2007) and Bock (1997).

## 9.6 Key Points and Further Issues

**Note:**

✓ We have introduced important point estimates for the parameters of a normal and a binomial distribution:

$$\bar{x} \text{ for } \mu, \qquad S^2 \text{ for } \sigma^2, \qquad \bar{x} \text{ for } p.$$

In general, the choice of these point estimates is not arbitrary but follows some principles of statistical inference such as maximum likelihood estimation, or least squares estimation (introduced in Chap. 11).

✓ The maximum likelihood estimator is usually consistent, asymptotically unbiased, asymptotically normally distributed, and asymptotically efficient.

✓ The validity of all results in this chapter depends on the assumption that the data is complete and has no missing values. Incomplete data may yield different conclusions.

✓ A confidence interval is defined in terms of upper and lower confidence limits and covers the true target parameter with probability $1 - \alpha$. Confidence intervals are often constructed as follows:

$$\text{point estimate} \pm \text{quantile} \cdot \underbrace{\sqrt{\text{variance of point estimate}}}_{\text{standard error}} .$$

✓ More detailed introductions to inference are presented in Casella and Berger (2002) and Young and Smith (2005).

## 9.7 Exercises

*Exercise 9.1* Consider an i.i.d. sample of size $n$ from a $\text{Po}(\lambda)$ distributed random variable $X$.

(a) Determine the maximum likelihood estimate for $\lambda$.
(b) What does the log-likelihood function look like for the following realizations: $x_1 = 4, x_2 = 3, x_3 = 8, x_4 = 6, x_5 = 6$? Plot the function using *R*. Hint: The `curve` command can be used to plot functions.
(c) Use the Neyman–Fisher Factorization Theorem to argue that the maximum likelihood estimate obtained in (a) is a sufficient statistic for $\lambda$.

*Exercise 9.2* Consider an i.i.d. sample of size $n$ from a $N(\mu, \sigma^2)$ distributed random variable $X$.

(a) Determine the maximum likelihood estimator for $\mu$ under the assumption that $\sigma^2 = 1$.
(b) Now determine the maximum likelihood estimator for $\mu$ for an arbitrary $\sigma^2$.
(c) What is the maximum likelihood estimate for $\sigma^2$?

*Exercise 9.3* Let $X_1, X_2, \ldots, X_n$ be $n$ i.i.d. random variables which follow a uniform distribution, $U(0, \theta)$. Write down the likelihood function and argue, without differentiating the function, what the maximum likelihood estimate of $\theta$ is.

*Exercise 9.4* Let $X_1, X_2, \ldots, X_n$ be $n$ i.i.d. random variables which follow an exponential distribution. An intelligent statistician proposes to use the following two estimators to estimate $\mu = 1/\lambda$:

(i) $T_n(X) = nX_{\min}$ with $X_{\min} = \min(X_1, \ldots, X_n)$ and $X_{\min} \sim \text{Exp}(n\lambda)$,
(ii) $V_n(X) = n^{-1} \sum_{i=1}^{n} X_i$.

(a) Are both $T_n(X)$ and $V_n(X)$ (asymptotically) unbiased for $\mu$?
(b) Calculate the mean squared error of both estimators. Which estimator is more efficient?
(c) Is $V_n(X)$ MSE consistent, weakly consistent, both, or not consistent at all?

*Exercise 9.5* A national park in Namibia determines the weight (in kg) of a sample of common eland antelopes:

450 730 700 600 620 660 850 520 490 670 700 820
910 770 760 620 550 520 590 490 620 660 940 790

Calculate

(a) the point estimate of $\mu$ and $\sigma^2$ and
(b) the confidence interval for $\mu$ ($\alpha = 0.05$).

under the assumption that the weight is normally distributed.

(c) Use *R* to reproduce the results from (b).

*Exercise 9.6* We are interested in the heights of the players of the two basketball teams "Brose Baskets Bamberg" and "Bayer Giants Leverkusen" as well as the football team "SV Werder Bremen". The following summary statistics are given:

| | $N$ | Minimum | Maximum | Mean | Std. dev. |
|---|---|---|---|---|---|
| Bamberg | 16 | 185 | 211 | 199.06 | 7.047 |
| Leverkusen | 14 | 175 | 210 | 196.00 | 9.782 |
| Bremen | 23 | 178 | 195 | 187.52 | 5.239 |

Calculate a 95 % confidence interval for $\mu$ for all three teams and interpret the results.

*Exercise 9.7* A married couple tosses a coin after each dinner to determine who has to wash the dishes. If the coin shows "head", then the husband has to wash the dishes, and if the coin shows "tails", then the wife has to wash the dishes. After 98 dinners, the wife notes that the coin has shown head 59 times.

(a) Estimate the probability that the wife has to wash the dishes.
(b) Calculate and interpret the 95 % confidence interval for $p$.
(c) How many dinners are needed to estimate the true probability for the coin showing "head" with a precision of $\pm 0.5$ % under the assumption that the coin is fair?

*Exercise 9.8* Suppose 93 out of 104 pupils have passed the final examination at a certain school.

(a) Calculate a 95 % confidence interval for the probability of failing the examination both by manual calculations and by using *R*, and compare the results.
(b) At county level 3.2 % of pupils failed the examination. Are the school's pupils worse than those in the whole county?

*Exercise 9.9* To estimate the audience rate for several TV stations, 3000 households are asked to allow a device, which records which TV station is watched, to be installed on their TVs. 2500 agreed to participate. Assume it is of interest to estimate the probability of someone switching on the TV and watching the show "Germany's next top model".

(a) What is the precision with which the probability can be estimated?
(b) What source of bias could potentially influence the estimates?

*Exercise 9.10* An Olympic decathlon athlete is interested in his performance compared with the performance of other athletes. He is a good runner and interested in his 100 m results compared with those of other athletes.

(a) He uses the decathlon data from this book (Appendix A.2) to come up with $\hat{\sigma} = s = 0.233$. What sample size does he need to calculate a 95 % confidence interval for the mean running time which is precise to $\pm 0.1$ s?
(b) Calculate a 95 % confidence interval for the mean running time ($\bar{x} = 10.93$) of the 30 athletes captured in the data set in Chap. A.2. Interpret the width of this interval compared with the width determined in a).
(c) The runner's own best time is 10.86 s. He wants to be among the best 10 % of all athletes. Calculate an appropriate confidence interval to compare his time with the 10 % best times.

*Exercise 9.11* Consider the pizza delivery data described in Chap. A.4. We distinguish between pizzas delivered on time (i.e. in less than 30 min) and not delivered on time (i.e. in more than 30 min). The contingency table for delivery time and operator looks as follows:

| | Operator | | Total |
|---|---|---|---|
| | Laura | Melissa | |
| <30 min | 163 | 151 | 314 |
| ≥30 min | 475 | 477 | 952 |
| Total | 638 | 628 | 1266 |

(a) Calculate and interpret the odds ratio and its 95 % confidence interval.
(b) Reproduce the results from (a) using *R*.

→ Solutions to all exercises in this chapter can be found on p. 384